

Simultaneous Determination of Water Constituent Concentrations from Airborne Imaging Spectrometer Data and Partial Least Squares

Allan Aasbjerg Nielsen

IMM, Informatics and Mathematical Modelling, Building 321

Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

e-mail aa@imm.dtu.dk, internet www.imm.dtu.dk/~aa

Abstract

This paper deals with a physically based bio-optical model for the simultaneous determination of concentrations of water constituents chlorophyll-a (CHL), total suspended matter (TSM), and coloured, dissolved organic matter (CDOM). The model is based on the dependency of absorption and backscatter coefficients on the constituent concentrations. A fundamental remote sensing equation is rewritten to a regression model which is used to estimate the concentrations of CHL and TSM. A weighted regression analysis is performed and the weights are determined as the weights chosen in a partial least squares (PLS) or canonical covariance analysis of in situ measurements of CHL, TSM and CDOM with in situ spectra in the 400-730 nm region. Resampled weights are used with geometrically corrected and calibrated airborne optical *casi* data to produce maps of jointly estimated CHL and TSM contents. Also, considerations to give a better understanding of the PLS technique than offered by the NIPALS algorithm are given and canonical correlations analysis is briefly described.

1 Introduction

Based on [1] we suggest a model for the joint estimation of concentrations of water constituents chlorophyll-a (CHL), total suspended matter (TSM), and coloured, dissolved organic matter (CDOM) [2]. Here, we estimate CHL and TSM only but the model can be used to estimate CDOM also. Although CDOM is not estimated its spectral behaviour is accounted for in the joint estimation of CHL and TSM.

Below the following symbols are used:

- $R(0-)$ is the irradiance reflected just below the water surface,
- a is the absorption coefficient,
- b_b is the backscatter coefficient,
- f is a proportionality constant (here $f = 0.50$),
- CHL is the concentration of chlorophyll-a,
- TSM is the concentration of total suspended matter,
- TSM_{tr} is the concentration of the part of TSM that is unrelated to CHL,
- TSM_{ph} is the concentration of the part of TSM that is related to CHL,
- f_c is a proportionality constant (here $f_c = 0.06$),
- a_w is the absorption coefficient for water,

- a_{tsmph}^* is the specific absorption coefficient for TSM_{ph},
- a_{chl}^* is the specific absorption coefficient for CHL,
- a_{cdom} is the absorption coefficient for CDOM,
- b_{bw} is the backscatter coefficient for water,
- b_{bstmtr}^* is the specific backscatter coefficient for TSM_{tr}.

2 A Bio-optical Model

A fundamental remote sensing equation, (1) below, and equations relating inherent optical properties (IOPs: a_w , a_{tsmph}^* , a_{chl}^* , a_{cdom} , b_{bw} and b_{bstmtr}^*) and the absorption and backscatter coefficients a and b_b

$$R(0-) = f \frac{b_b}{a + b_b} \quad (1)$$

$$a = a_w + a_{tsmph}^* \cdot \text{TSM}_{ph} + a_{chl}^* \cdot \text{CHL} + a_{cdom} \quad (2)$$

$$b_b = b_{bw} + b_{bstmtr}^* \cdot \text{TSM}_{tr} \quad (3)$$

$$\text{TSM} = \text{TSM}_{tr} + \text{TSM}_{ph} = \text{TSM}_{tr} + f_c \cdot \text{CHL} \quad (4)$$

are rewritten using the auxiliary variable $x = 1 - f/R(0-)$ leading to $a + b_b \cdot x = 0$ or

$$\begin{aligned} -(a_w + a_{cdom} + b_{bw} \cdot x) = & \\ (a_{chl}^* + f_c \cdot (a_{tsmph}^* - b_{bstmtr}^* \cdot x)) \cdot \text{CHL} & \\ + b_{bstmtr}^* \cdot x \cdot \text{TSM}. & \end{aligned} \quad (5)$$

The IOPs and $R(0-)$ and therefore also x are wavelength dependent and the above equation is valid for all m wavelengths available. The equations for all wavelengths can be written in vector and matrix notation as

$$\mathbf{z} = \mathbf{A}\boldsymbol{\beta} + \mathbf{n} \quad (6)$$

where

- \mathbf{z} ($m \times 1$) contains $-(a_w + a_{cdom} + b_{bw} \cdot x)$ for all m wavelengths available,
- \mathbf{A} ($m \times 2$) in the first column contains $a_{chl}^* + f_c \cdot (a_{tsmph}^* - b_{bstmtr}^* \cdot x)$ and in the second column contains $b_{bstmtr}^* \cdot x$ for all m wavelengths available, and
- $\boldsymbol{\beta}$ (2×1) is $[\text{CHL } \text{TSM}]^T$ to be estimated;
- \mathbf{n} ($m \times 1$) are the residuals.

In [1] a slightly different model is used for two wavelengths (675 nm and 705 nm) to solve for the two unknowns CHL and TSM_{tr}. Here, we apply more wavelengths in a (weighted) regression analysis to obtain the desired estimates.

3 Methods

Earlier work on the model in [1] based on ordinary least squares (OLS) regression showed that results depended heavily on the wavelengths chosen. A weighted least squares (WLS) regression is therefore performed. To find the weights to be applied partial least squares (PLS) or canonical covariance analysis is chosen.

3.1 PLS regression

In partial least squares, PLS, we consider two multivariate sets \mathbf{X} ($m \times 1$) and \mathbf{Y} ($k \times 1$) with \mathbf{Y} considered as response variables. \mathbf{X} often contains (many) spectral variables, $m > k$ or $m \gg k$. In this notation \mathbf{X} and \mathbf{Y} are vector random variables, one for each observation. Often the number of observations n is small. PLS is normally described as a black-box method (as the so-called NIPALS algorithm). A thorough understanding of PLS is prevented by this description. A more “multivariate-statistics-oriented” description is given here.

We want to maximise the covariance R between linear combinations $t = \mathbf{w}^T \mathbf{X}$ of \mathbf{X} and $u = \mathbf{c}^T \mathbf{Y}$ of \mathbf{Y} , $R = \text{Cov}\{t, u\} = \mathbf{w}^T \boldsymbol{\Sigma}_{12} \mathbf{c}$, under the constraints $\mathbf{w}^T \mathbf{w} = \mathbf{c}^T \mathbf{c} = 1$. $\boldsymbol{\Sigma}_{12}$ ($m \times k$) is the covariance between \mathbf{X} and \mathbf{Y} ($\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T$). To maximise R we use a Lagrange multiplier technique and introduce $F = R - (\lambda_1/2)(\mathbf{w}^T \mathbf{w} - 1) - (\lambda_2/2)(\mathbf{c}^T \mathbf{c} - 1)$ which we maximise without constraints. This is done by setting $\partial F / \partial \mathbf{w} = \partial F / \partial \mathbf{c} = \mathbf{0}$ leading to

$$\boldsymbol{\Sigma}_{12} \mathbf{c} = \lambda_1 \mathbf{w} \quad (7)$$

$$\boldsymbol{\Sigma}_{21} \mathbf{w} = \lambda_2 \mathbf{c}. \quad (8)$$

(Setting $\partial F / \partial \lambda_i = 0$ merely reproduce the constraints.) Multiplying (7) by \mathbf{w}^T and (8) by \mathbf{c}^T we see that $\lambda_1 = \lambda_2 = R$ and by substituting \mathbf{w} from (7) into (8) and \mathbf{c} from (8) into (7) we get

$$\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{21} \mathbf{w} = R^2 \mathbf{w} \quad (9)$$

$$\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{12} \mathbf{c} = R^2 \mathbf{c} \quad (10)$$

[3, 4], i.e., we find the desired projections for \mathbf{X} by considering the conjugate eigenvectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ corresponding to the eigenvalues $R_1^2 \geq \dots \geq R_k^2$ of $\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{21}$. Similarly, we may find the desired projections for \mathbf{Y} by considering the conjugate eigenvectors $\mathbf{b}_1, \dots, \mathbf{b}_k$ of $\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{12}$ corresponding to the same eigenvalues R_i^2 . (As the solutions \mathbf{w} and \mathbf{c} are interrelated by (7) and (8) we only need find one of them.) [4] also hints a way to perform multiset or multiblock PLS. In this situation \mathbf{X} naturally splits into several sets of blocks.

If $k = 1$, i.e., the response is univariate ($\mathbf{Y} = Y$) an eigensolution is not needed. In this case \mathbf{c} is the scalar one ($\mathbf{c} = c = 1$) and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\sigma}_{12}$ (an $m \times 1$ vector), $\mathbf{w} = \boldsymbol{\sigma}_{12} / \sqrt{\boldsymbol{\sigma}_{12}^T \boldsymbol{\sigma}_{12}}$, and $R = \mathbf{w}^T \boldsymbol{\sigma}_{12}$.

This is very similar to canonical correlations analysis [5, 6, 7, 8], in which the correlation $\rho = \text{Corr}\{\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{Y}\} = \mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b} / \sqrt{(\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a})(\mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b})}$ is maximised. To do this we set $\partial \rho / \partial \mathbf{a} = \partial \rho / \partial \mathbf{b} = \mathbf{0}$ and get

$$\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a} \boldsymbol{\Sigma}_{12} \mathbf{b} = R \boldsymbol{\Sigma}_{11} \mathbf{a} \quad (11)$$

$$\mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b} \boldsymbol{\Sigma}_{21} \mathbf{a} = R \boldsymbol{\Sigma}_{22} \mathbf{b}. \quad (12)$$

Without loss of generality we choose \mathbf{a} and \mathbf{b} so that $\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a} = \mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b} = 1$, i.e., the new variables called canonical variates have unit variance, which leads to

$$\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \mathbf{a} = \rho^2 \boldsymbol{\Sigma}_{11} \mathbf{a} \quad (13)$$

$$\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \mathbf{b} = \rho^2 \boldsymbol{\Sigma}_{22} \mathbf{b}, \quad (14)$$

i.e., inversion of $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ is needed. This is not feasible when the number of observations is small.

In PLS only the first pair of canonical variates (or latent variables) t and u corresponding to the eigensolutions with the largest eigenvalue are calculated and the response u is regressed on

the predictor t : $u = bt + e$. Loadings are defined as $\mathbf{p} = \Sigma_{11}\mathbf{w}/\mathbf{w}^T\Sigma_{11}\mathbf{w}$ (X-loadings) and $\mathbf{q} = \Sigma_{22}\mathbf{c}/\mathbf{c}^T\Sigma_{22}\mathbf{c}$ (Y-loadings). If more information is present in the residuals e these are subtracted from the original response variables (i.e., \mathbf{Y} is replaced by $\mathbf{Y} - bt\mathbf{c}$), the predictor variables are projected onto a subspace orthogonal to the solution found (i.e., \mathbf{X} is replaced by $\mathbf{X} - t\mathbf{p}$ or $\mathbf{X} - t\mathbf{w}$, the former is normally chosen) and the procedure is repeated on the new \mathbf{X} and \mathbf{Y} , see also [9, 10, 11, 12].

Here, \mathbf{X} contains in situ spectra ($R(0-)$) covering from 400 nm to 730 nm in 2 nm intervals ($m = 166$). \mathbf{Y} is in situ [CHL TSM CDOM] T ($k = 3$). 43 observations from various coastal regions in Denmark were available. Figure 1 shows the weights and loadings obtained from a PLS (or canonical covariance) analysis with one latent variable. Also, jackknife weights obtained by leaving out one observation at a time are shown. Since these jackknife weights seem stable all 43 observations are used. Very low weights for 400-402 nm and negative weights for 404-536 nm are set to 0 in the WLS regression.

3.2 WLS regression

The ordinary least squares (OLS) regression solution to (6) obtained by minimising the sum of squared residuals, i.e., by setting $\partial(\mathbf{n}^T\mathbf{n})/\partial\boldsymbol{\beta} = \mathbf{0}$ is

$$\mathbf{A}^T\mathbf{A}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{A}^T\mathbf{z} \quad (15)$$

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{z}. \quad (16)$$

The weighted least squares (WLS) regression solution obtained by minimising the sum of weighted squared residuals, i.e., by setting $\partial(\mathbf{n}^T\mathbf{W}\mathbf{n})/\partial\boldsymbol{\beta} = \mathbf{0}$ is

$$\mathbf{A}^T\mathbf{W}\mathbf{A}\hat{\boldsymbol{\beta}}_{\text{WLS}} = \mathbf{A}^T\mathbf{W}\mathbf{z} \quad (17)$$

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{z} \quad (18)$$

where \mathbf{W} is a diagonal matrix with weights for each wavelength available (in both cases provided of course that the inverse matrices exist). The weights are obtained using partial least squares (or canonical covariance) analysis as described above. Since in this case both \mathbf{z} and \mathbf{A} are based on measured quantities, orthogonal regression could be considered. This idea is not pursued further here.

4 Results with Airborne Data

With geometrically corrected and calibrated airborne *casi* (compact airborne spectrographic imager, [13]) data shown in Figure 2, we apply WLS regression in model (5) with 12 of the 19 recorded wavelengths. The *casi* data are recorded over Århus Bugt, Denmark, on 7 August 1999, pixels are 4×4 m², and the image size is 8,892 rows by 787 columns. In the figures this one flight line is shown as three columns of data. The weights applied are nearest neighbour versions of the weights chosen as described above. The IOPs applied are nearest neighbour versions of region specific in situ IOPs. The WLS regression gives simultaneous estimates for CHL (stretched linearly between 0 and 3 $\mu\text{g/l}$) and TSM (stretched linearly between 0 and 0.5 mg/l) shown in Figures 3 and 4. In both cases the colour scale goes from blue over cyan-green-yellow to red. Simultaneously recorded in situ measurements (5 observations) of CHL lie in the interval 1.6–2.8 $\mu\text{g/l}$. For TSM in situ measurements lie in the interval 0.7–1.7 mg/l.

5 Conclusions

Based on inherent optical properties established from in situ sampling and in situ irradiance measurements a framework for physically based simultaneous estimation of water constituent concentrations by means of weighted least squares regression is described. The weights applied in the analysis are obtained from a one latent variable partial least squares model of 43 observations of in situ measurements of water constituent concentrations and the spectral response in the 400-730 nm wavelength region. The estimates obtained for chlorophyll-a lie in the same interval as the in situ measurements whereas the estimated values for total suspended matter seem low.

Acknowledgments

This work was carried out as a part of the DECO project [2] headed by Dr. Bo Riemann, the National Environmental Research Institute (NERI), and funded by the Danish National Research Councils. The in situ IOPs were provided by Dr. Stiig Markager, Peter Stæhr, both NERI, and Dr. Peter V. Jørgensen, the Danish Meteorological Institute, [14]. PVJ also performed the *casi* preprocessing.

References

- [1] H. J. Hoogenboom, A. G. Dekker, and J. F. de Haan, "Retrieval of chlorophyll and suspended matter from imaging spectrometer data by matrix inversion," *Canadian Journal of Remote Sensing*, vol. 24, no. 2, pp. 144–152, 1998.
- [2] DECO, "Danish environmental monitoring of coastal waters (DECO), final report," Tech. Rep., National Environmental Research Institute et al., Denmark, 2001.
- [3] M. Borga, *Learning Multidimensional Signal Processing*, Ph.D. thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1998.
- [4] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data," *Accepted for IEEE Transactions on Image Processing*, 2001.
- [5] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. XXVIII, pp. 321–377, 1936.
- [6] W. W. Cooley and P. R. Lohnes, *Multivariate Data Analysis*, John Wiley and Sons, New York, 1971.
- [7] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, second edition, 1984.
- [8] A. A. Nielsen, *Analysis of Regularly and Irregularly Sampled Spatial, Multivariate, and Multi-temporal Data*, Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, 1994, Internet <http://www.imm.dtu.dk/~aa/phd/>.
- [9] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM Journal of Scientific Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.
- [10] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.
- [11] A. Höskuldsson, "PLS regression methods," *Journal of Chemometrics*, vol. 2, pp. 211–228, 1986.
- [12] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.
- [13] *casi*, "<http://www.itres.com/>," ITRES Research Limited, Calgary, Alberta, Canada.
- [14] N. K. Højerslev and P. V. Jørgensen, "Modelling of key bio-optical parameters including spectral light backscattering absorption coefficients using in-situ measurements of spectral up- and downwelling irradiances and light backscattering," *Submitted*, 2001.

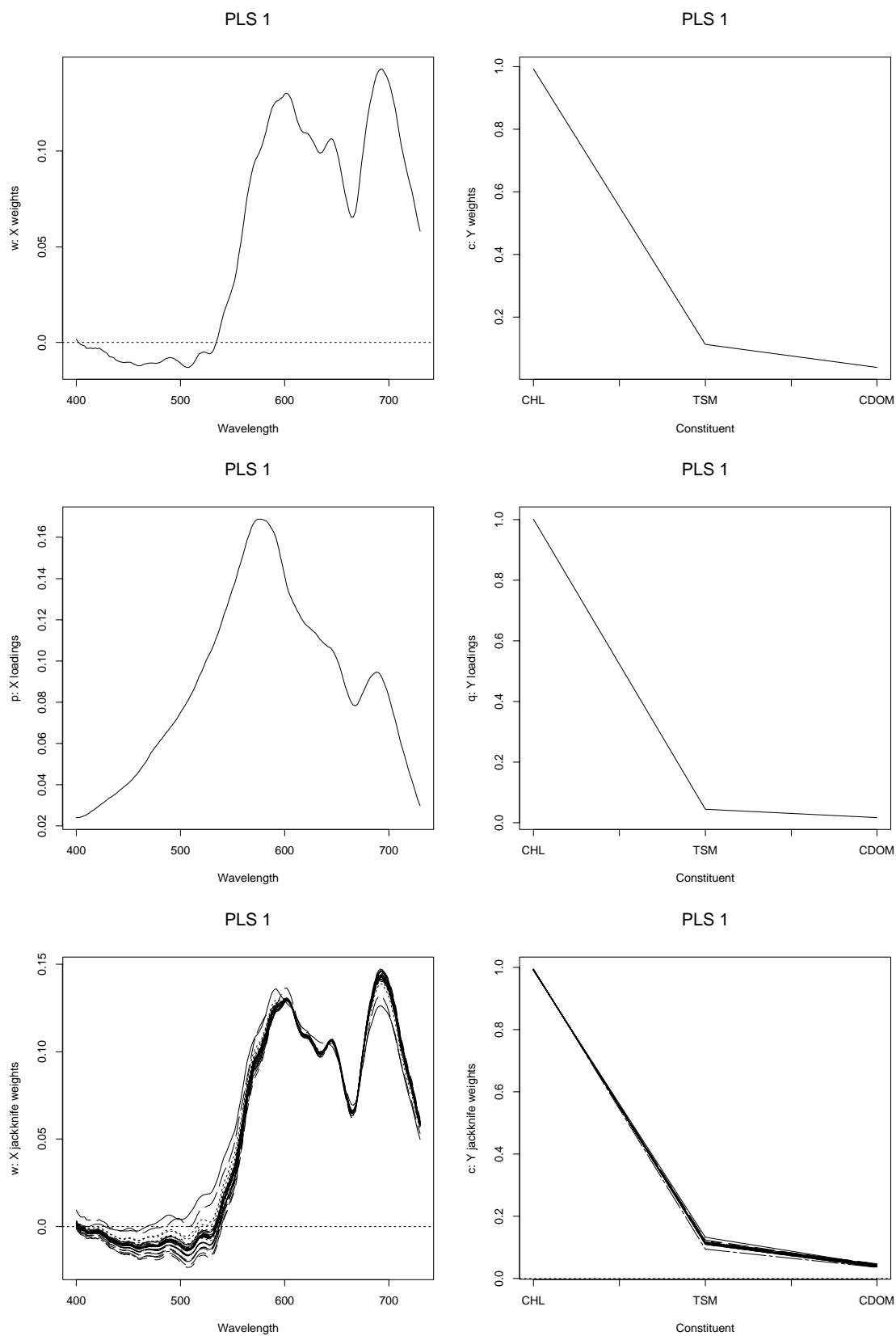


Figure 1: Weights, loadings and jackknife weights from PLS with one latent variable



Figure 2: Simulated natural colour RGB plot of *casi* data; the flight line is broken into three strips each 12,000 m long and 3,148 m across

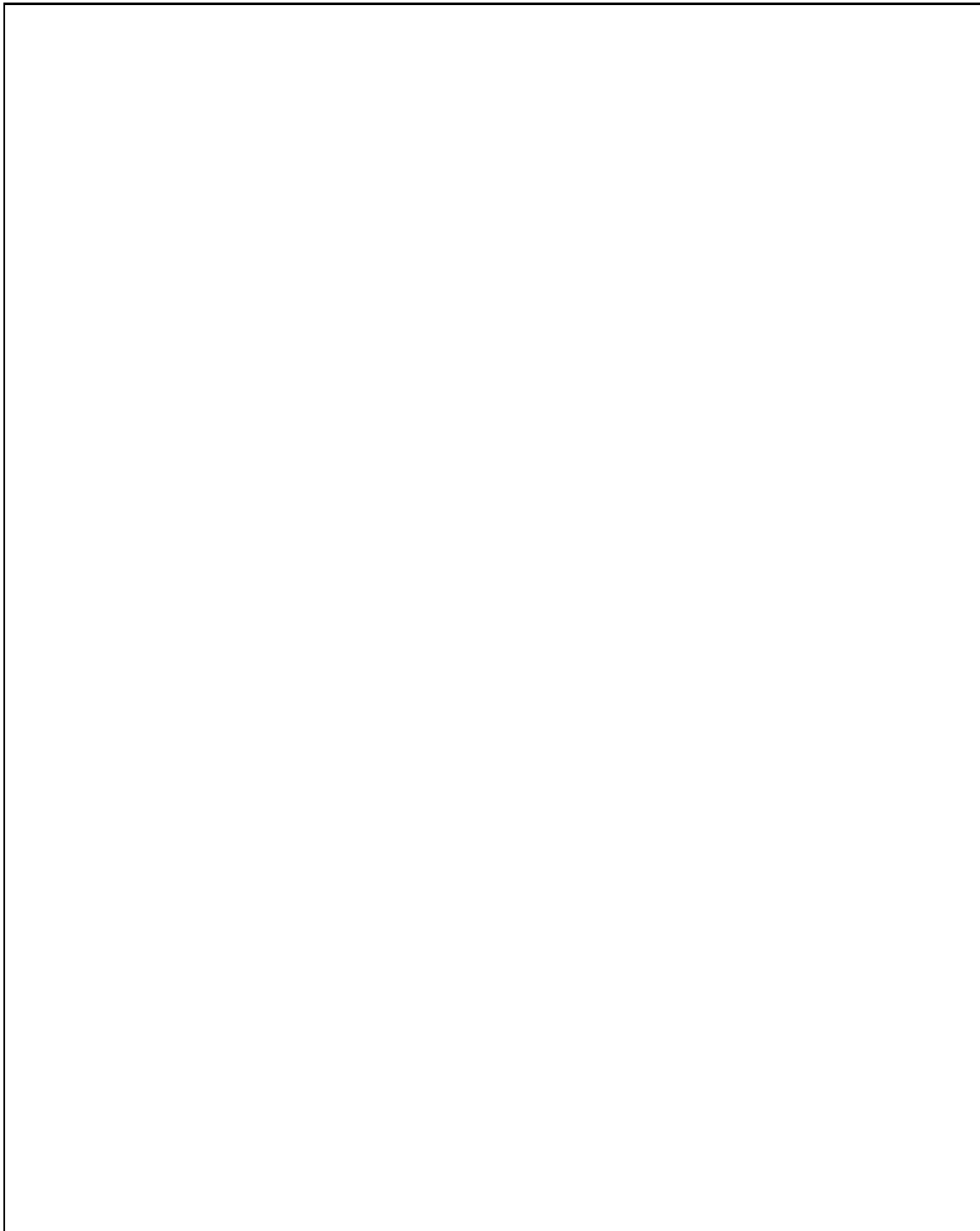


Figure 3: CHL estimate from *casi* data, stretch is linear from 0 to 3 $\mu\text{g/l}$, colour scale is from blue over cyan-green-yellow to red; the flight line is broken into three strips each 12,000 m long and 3,148 m across



Figure 4: TSM estimate from *casi* data, stretch is linear from 0 to 0.5 mg/l, colour scale is from blue over cyan-green-yellow to red; the flight line is broken into three strips each 12,000 m long and 3,148 m across